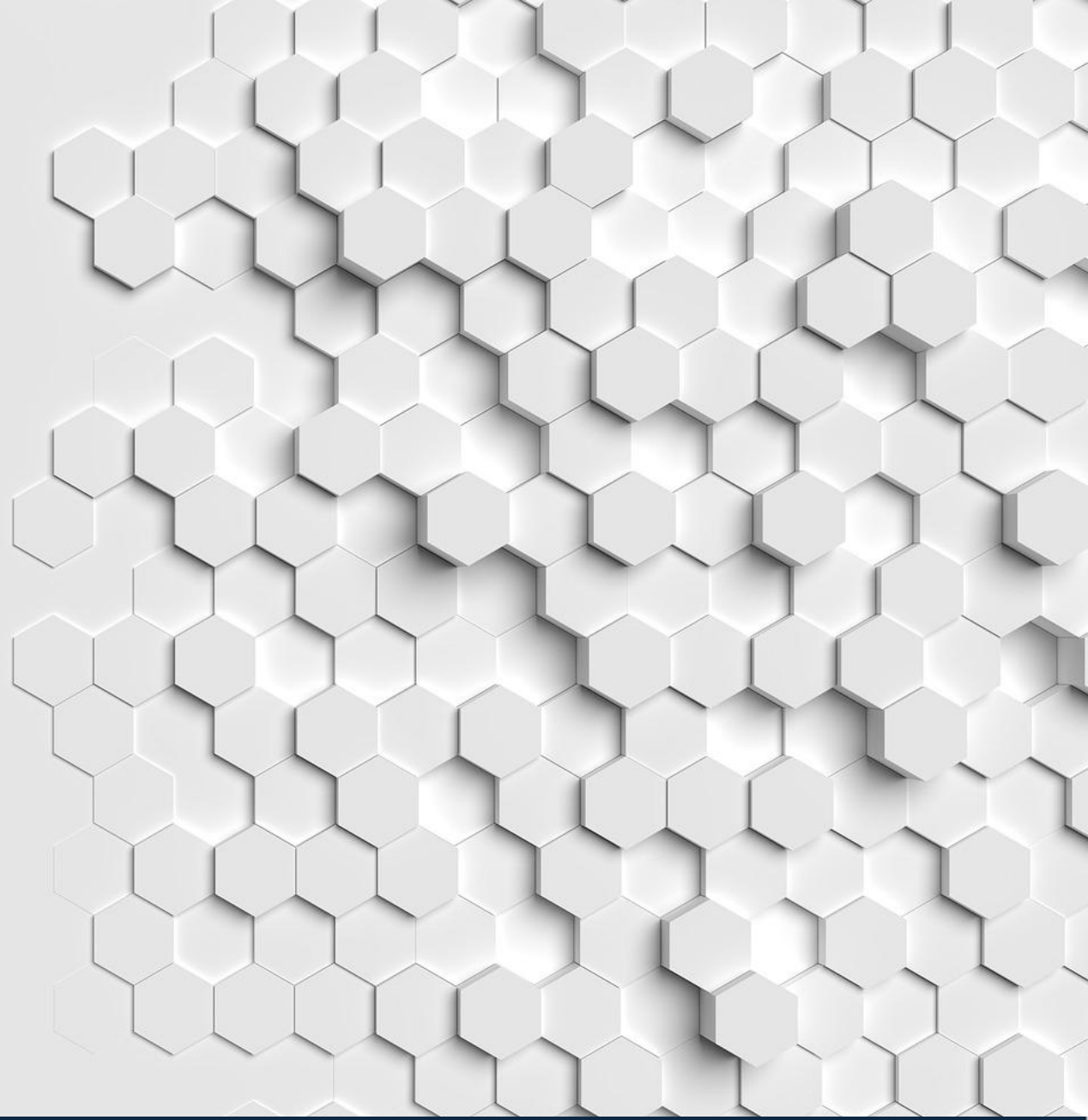
Biostatistics

Statistical computing
Luca Alberto Rizzo



CIDRA-AIPD
African Institute
for Professional
Development

Agenda

1. Introduction to R
2. Penguins dataset
3. Histogram, barchart & boxplots
4. Statistical tests for means and proportions
5. ANOVA
6. Conclusions

1 Introduction to R

Introduction to R: why learn R?

CIDRA-AIPD
African Institute
for Professional
Development

Lingua franca for statistical computing

Free and open source

Latest model/tech implemented

why learn

Robust visualization libraries

Used in industry & academia

Large and active community

Introduction to R: what is R?

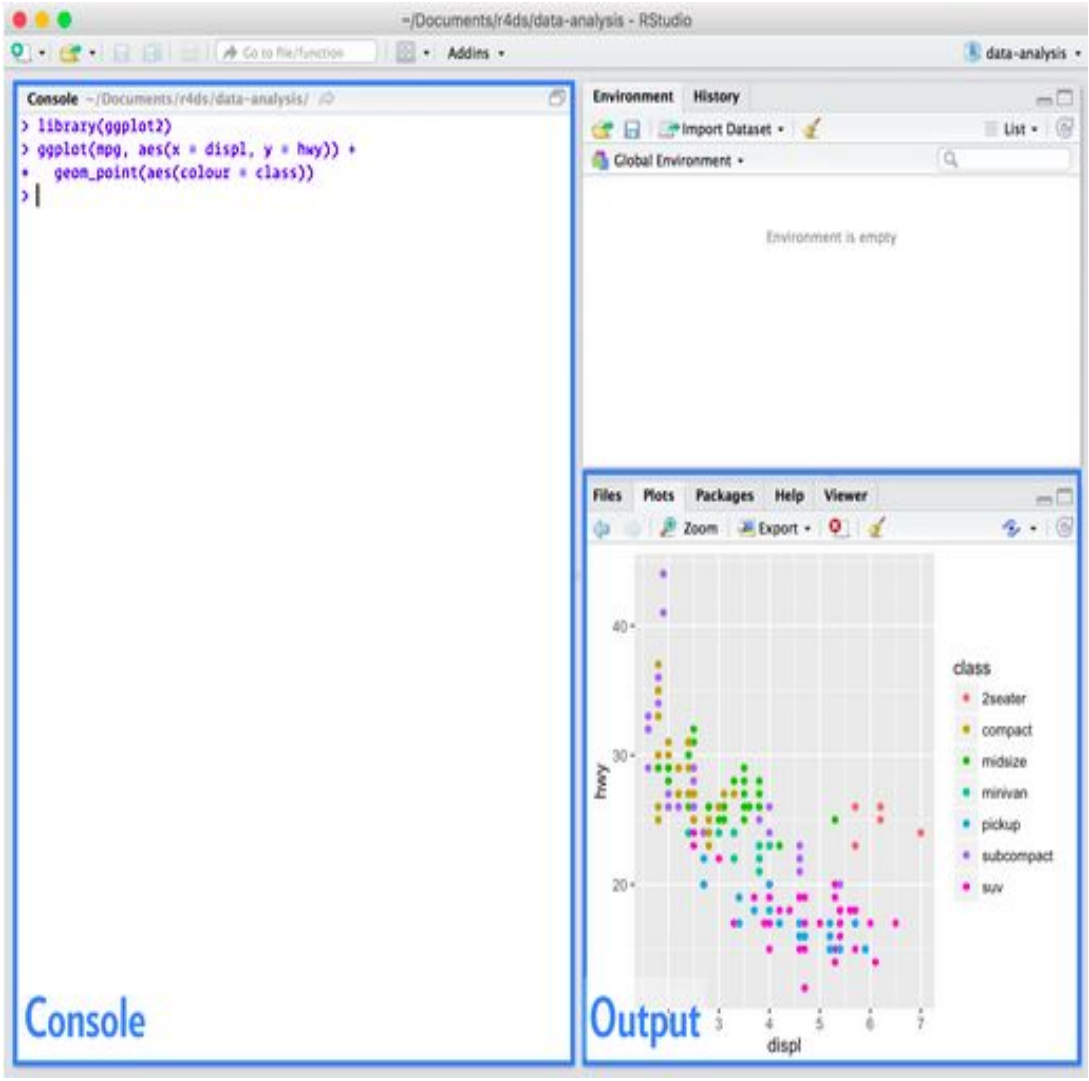
CIDRA-AIPD
African Institute
for Professional
Development

Introduction

- R is a language and environment for statistical computing and graphics
- R is developed and maintained by the R foundation
- R has >17,000 packages, with additional functions

Installation

- R can be installed easily via this link
- RStudio is an integrated development environment, or IDE, for R programming: It can be found here



Introduction to R: “hello world” and tidyverse

CIDRA-AIPD African Institute for Professional Development

1. Let’s start by greeting the whole world!

```
> print("hello world")
[1] "hello world"
```

2. Install tidyverse, a bundle of ~25 packages designed to help with data management, reproducibility and multi-level programming

```
> install.packages("tidyverse")
Installing package into ‘/home/lumaca/R/x86_64-pc-linux-gnu-library/4.2’
(as ‘lib’ is unspecified)
```

3. Let’s compute our **first mathematical operation in R** (mean)

```
> (1+2+3+4) / 4
[1] 2.5
```

4. Assign a vector and compute its mean with **the built-in R function**

```
> (x <- c(1,2,3,4))
[1] 1 2 3 4
```

```
> mean(x)
[1] 2.5
```

Take away:
Do not reinvent the wheel!

Introduction to R: data types

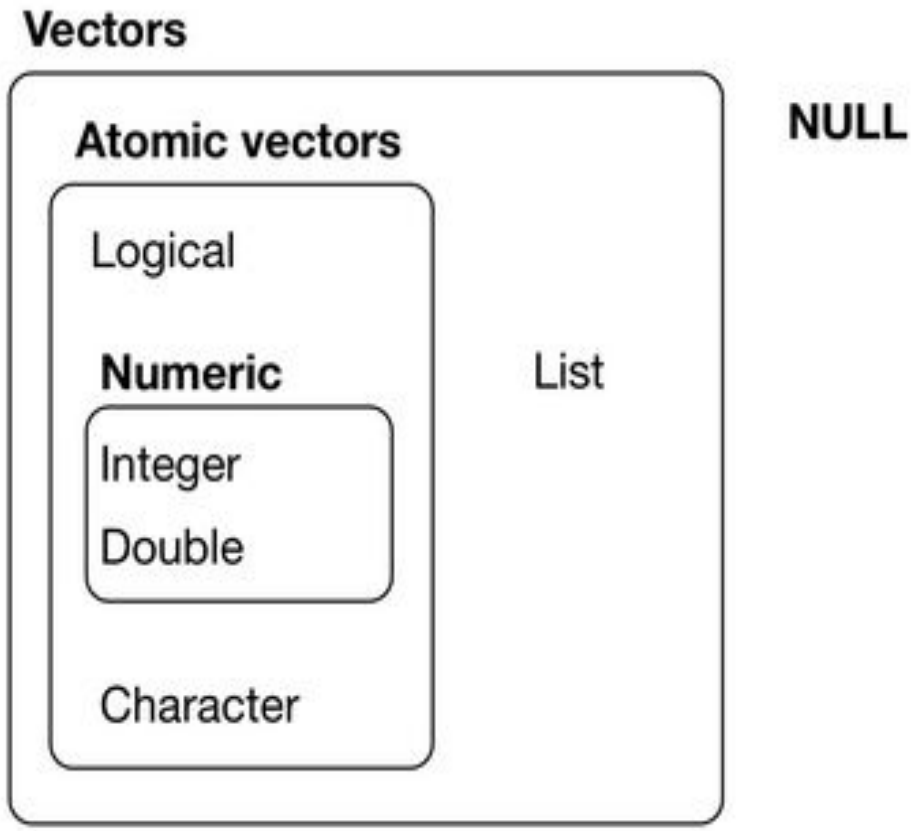
CIDRA-AIPD African Institute for Professional Development

Vectors and types in R

- **6 atomic (simple) vectors:** logical, integer, double, character, complex, raw
- Integers and doubles **are numeric**
- **c operator** builds non-atomic vectors
- **a list** is a heterogenous (recursive) vector

```
> typeof(1)
[1] "double"
> typeof(1L)
[1] "integer"
> typeof('a')
[1] "character"
```

```
> typeof(TRUE)
[1] "logical"
> typeof(NaN)
[1] "double"
> typeof(c(1,1L,NaN))
[1] "double"
```



Source: Figure 20.1 “R for Data Science”

Introduction to R: operators

CIDRA-AIPD African Institute for Professional Development

Operators in R

- the assignment operator in R is both:

  <- (**good practice**)

  = (**bad practice**)

- **==** checks equality between 2 elements

```
> (a <- 2)
[1] 2
```

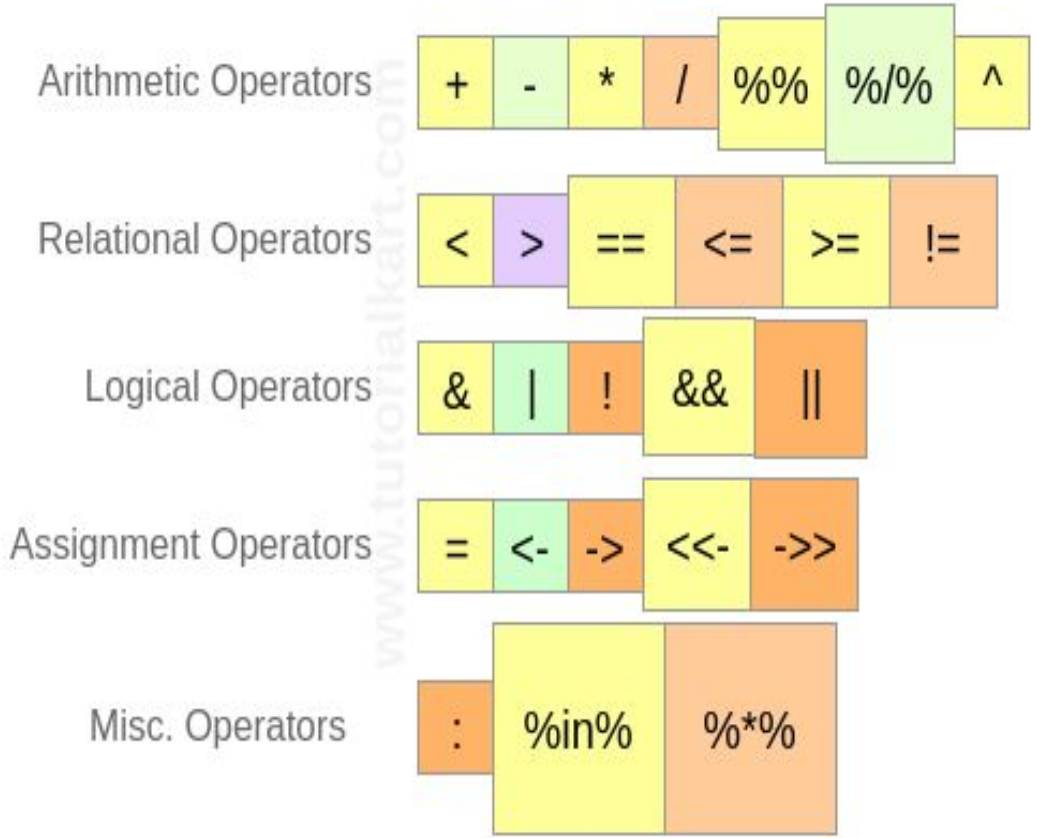
Assign with **<-**

```
> (b = 3)
[1] 3
```

Assign with **=**

```
> a == b
[1] FALSE
```

Checking equality with **==**



Souce: [tutorial kart website]

Introduction to R: dataframe & tibbles

CIDRA-AIPD African Institute for Professional Development

A dataframe is a **2D data structure in R**, a **special case of a list** which has each component **of equal length.**

```
> print(penguins)
# A tibble: 344 × 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex      year
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int> <fct>   <int>
 1 Adelie  Torgersen           39.1          18.7               181        3750 male     2007
 2 Adelie  Torgersen           39.5          17.4               186        3800 female   2007
 3 Adelie  Torgersen           40.3          18                 195        3250 female   2007
 4 Adelie  Torgersen           NA            NA                  NA          NA NA       2007
 5 Adelie  Torgersen           36.7          19.3               193        3450 female   2007
 6 Adelie  Torgersen           39.3          20.6               190        3650 male     2007
 7 Adelie  Torgersen           38.9          17.8               181        3625 female   2007
 8 Adelie  Torgersen           39.2          19.6               195        4675 male     2007
 9 Adelie  Torgersen           34.1          18.1               193        3475 NA       2007
10 Adelie  Torgersen           42            20.2               190        4250 NA       2007
# … with 334 more rows
```

This is a **tibble, a special version of a dataframe**
implemented in the tidyverse library

Introduction to R: dataframe & tibbles

CIDRA-AIPD African Institute for Professional Development

Summary prints useful information

```
> summary(penguins)
      species          island    bill_length_mm  bill_depth_mm   flipper_length_mm
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10   Min.   :172.0
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30   Median :197.0
                                 Mean   :43.92   Mean   :17.15   Mean   :200.9
                                 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0
                                 Max.   :59.60   Max.   :21.50   Max.   :231.0
                                 NA's   :2       NA's   :2       NA's   :2
  body_mass_g       sex           year
 Min.   :2700   female:165   Min.   :2007
 1st Qu.:3550   male  :168   1st Qu.:2007
 Median :4050   NA's  : 11   Median :2008
 Mean   :4202                Mean   :2008
 3rd Qu.:4750                3rd Qu.:2009
 Max.   :6300                Max.   :2009
 NA's   :2
```

Mutate adds columns

```
> mutate(penguins,
+          bill_ratio = bill_length_mm / bill_depth_mm)
# A tibble: 344 × 9
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex     year bill_ratio
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int> <fct>  <int>      <dbl>
 1 Adelie  Torgersen           39.1          18.7               181        3750 male    2007       2.09
 2 Adelie  Torgersen           39.5          17.4               186        3800 female  2007       2.27
 3 Adelie  Torgersen           40.3          18                 195        3250 female  2007       2.24
 4 Adelie  Torgersen           NA            NA                  NA          NA NA      2007      NA
 5 Adelie  Torgersen           36.7          19.3               193        3450 female  2007       1.90
 6 Adelie  Torgersen           39.3          20.6               190        3650 male    2007       1.91
 7 Adelie  Torgersen           38.9          17.8               181        3625 female  2007       2.19
 8 Adelie  Torgersen           39.2          19.6               195        4675 male    2007       2
 9 Adelie  Torgersen           34.1          18.1               193        3475 NA      2007       1.88
10 Adelie  Torgersen           42            20.2               190        4250 NA      2007       2.08
# … with 334 more rows
```

Filter and select

```
> penguins %>% select(species, island, body_mass_g) %>% filter(species == "Adelie")
# A tibble: 152 × 3
   species island    body_mass_g
   <fct>   <fct>           <int>
 1 Adelie  Torgersen        3750
 2 Adelie  Torgersen        3800
 3 Adelie  Torgersen        3250
 4 Adelie  Torgersen          NA
 5 Adelie  Torgersen        3450
 6 Adelie  Torgersen        3650
 7 Adelie  Torgersen        3625
 8 Adelie  Torgersen        4675
 9 Adelie  Torgersen        3475
10 Adelie  Torgersen        4250
# … with 142 more rows
```

Group by computes quantities per categorical variable

```
> group_by(penguins, species) %>%
+     summarise(avg_mass = mean(body_mass_g,  na.rm = TRUE))
# A tibble: 3 × 2
  species   avg_mass
  <fct>        <dbl>
1 Adelie       3701.
2 Chinstrap    3733.
3 Gentoo       5076.
```

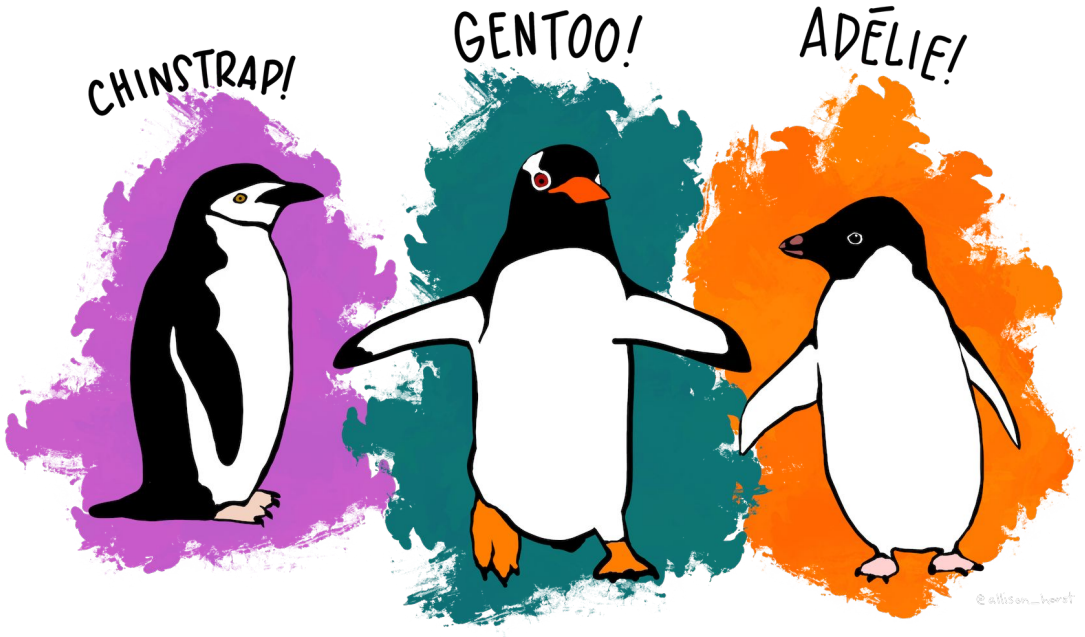
2 Penguins dataset

Penguins dataset: introduction

CIDRA-AIPD African Institute for Professional Development

Penguins dataset was collected by Dr. Kristen Gorman at the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

```
> summary(penguins)
      species          island    bill_length_mm  bill_depth_mm   flipper_length_mm
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10   Min.   :172.0
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30   Median :197.0
                                 Mean   :43.92   Mean   :17.15   Mean   :200.9
                                 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0
                                 Max.   :59.60   Max.   :21.50   Max.   :231.0
                                 NA's   :2       NA's   :2       NA's   :2
  body_mass_g       sex           year
 Min.   :2700   female:165   Min.   :2007
 1st Qu.:3550   male  :168   1st Qu.:2007
 Median :4050   NA's  : 11   Median :2008
 Mean   :4202                Mean   :2008
 3rd Qu.:4750                3rd Qu.:2009
 Max.   :6300                Max.   :2009
 NA's   :2
```



Artwork by @allison_hors

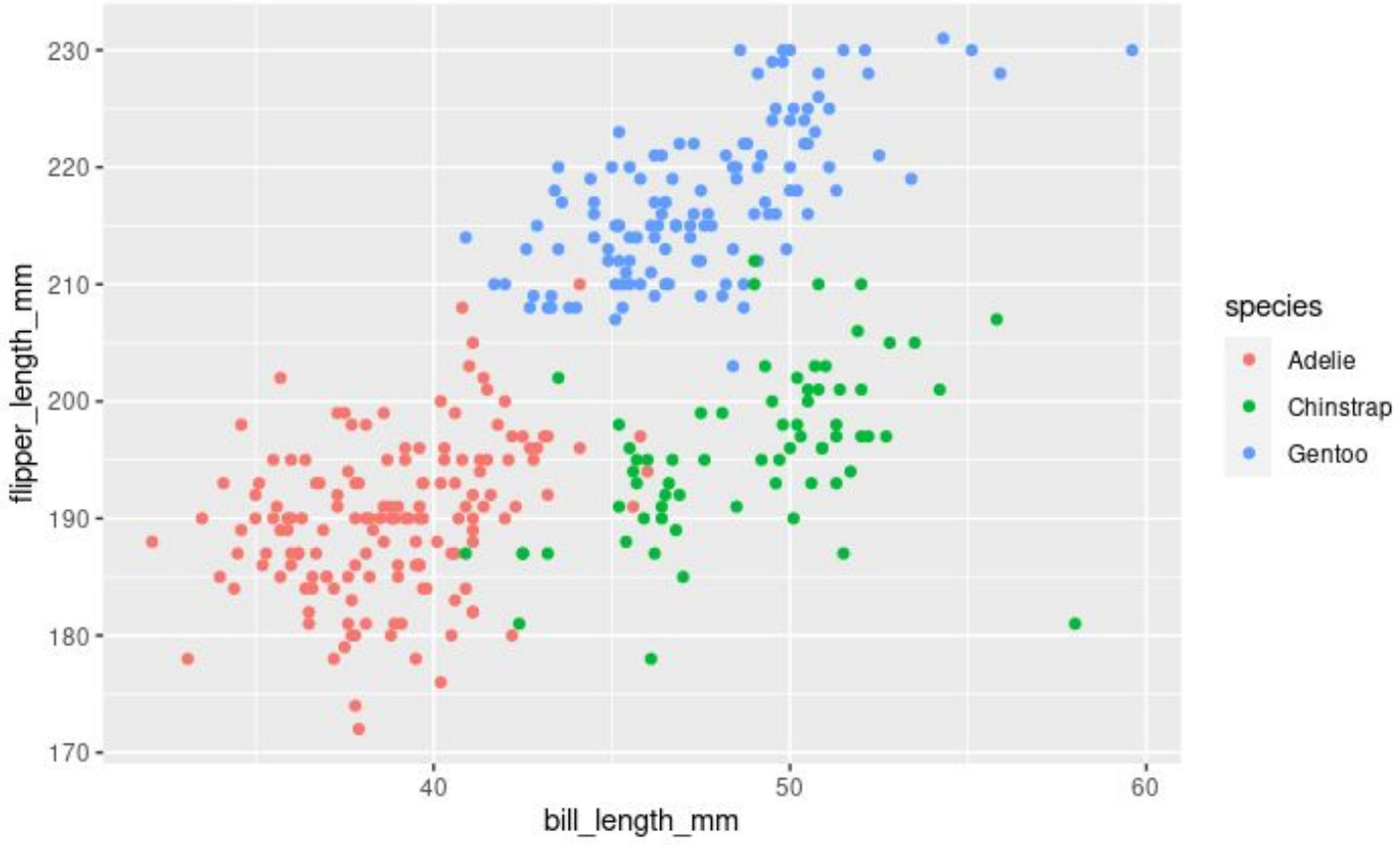
- 334 rows and 8 columns
- 3 species of penguins (Chinstrap, Gentoo, Adelie)
- 3 different islands (Biscoe, Dream, Torgersen)
- 3 factors (*species, islands, sex*), 2 doubles (*bill_lenght_mm, bill_depht_mm*) and 3 integers (*flipper_length_mm, body_mass_g, year*)

3 Histogram, barcharts, boxplots

Histogram, barcharts, boxplots: ggplot

"ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics."

```
> ggplot(penguins, aes(bill_length_mm, flipper_length_mm, colour = species )) + geom_point()
```

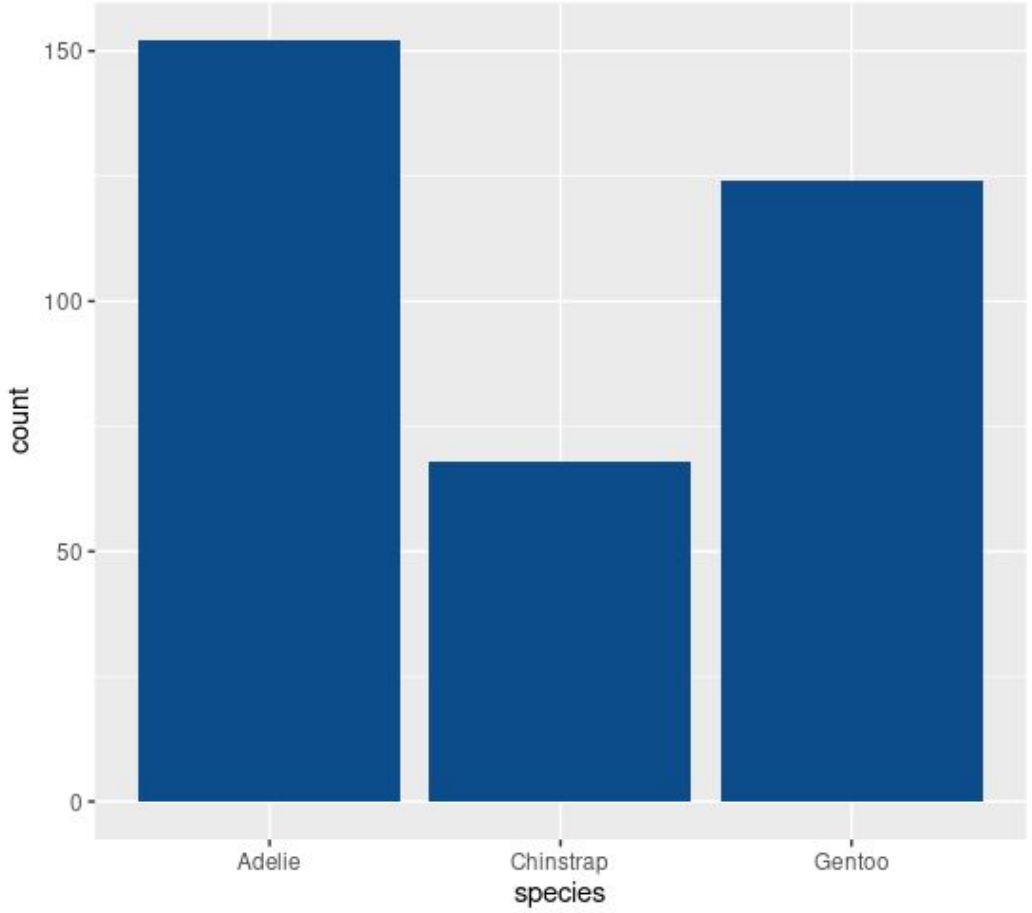


- ggplot(data = NULL, mapping = aes(), ...,) initializes a ggplot object
- aes(x,y,...) "Aesthetic mappings" describes how variables in the data are mapped to visual properties
- layers, like geom_point(), specify which kind of plot you want to produce
- ggplot cheasheet for more functionalities

Histogram, barcharts, boxplots: histogram

How many penguins of each species are there in the dataset?

```
> ggplot(penguins) + aes(x=species) + geom_bar(stat= "count", fill = "#0c4c8a")
```

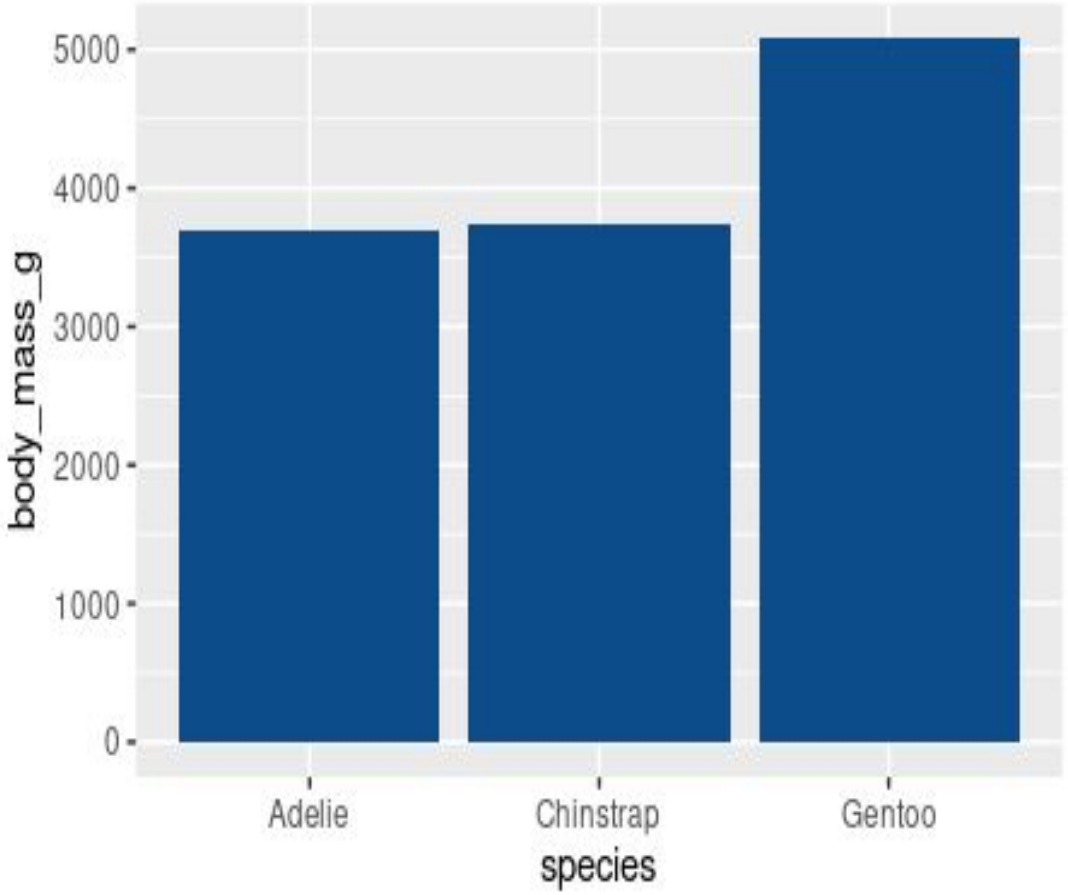


- geom_bar(mapping = NULL, data = NULL, stat = "count", …) is the layer which prints “bar charts”
- **if geom_bar(if stat= “count”,...)** a frequency histogram is plotted
- **if geom_bar(aes(y = (..count..)/sum(..count..)), ...),** a relative frequency histogram is plotted

Histogram, barcharts, boxplots: boxplots

What is the average mass of penguins for each specie?

```
> ggplot(data=penguins, aes(x=species, y=body_mass_g)) + geom_bar(stat = "summary", fun= "mean", fill = "#0c4c8a")
```



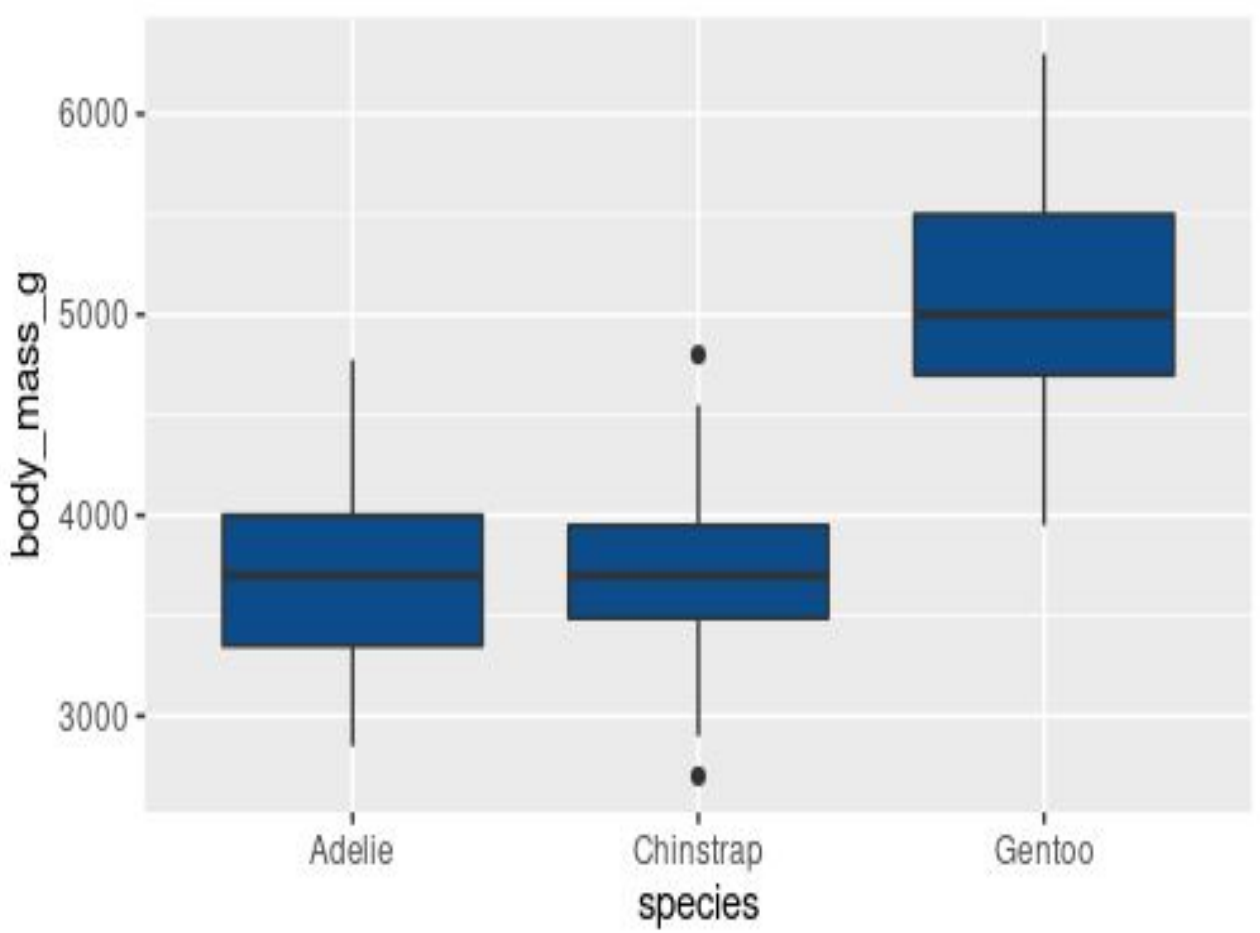
- **geom_bar(stat= “summary”,fun = “mean”)** we specified that we use a summary statistic, in particular the mean

Histogram, barcharts, boxplots: boxplots

CIDRA-AIPD African Institute for Professional Development

Can we have more information about the distribution of mass among species?

```
> ggplot(data = penguins, mapping = aes(x = species, y = body_mass_g)) + geom_boxplot()
```



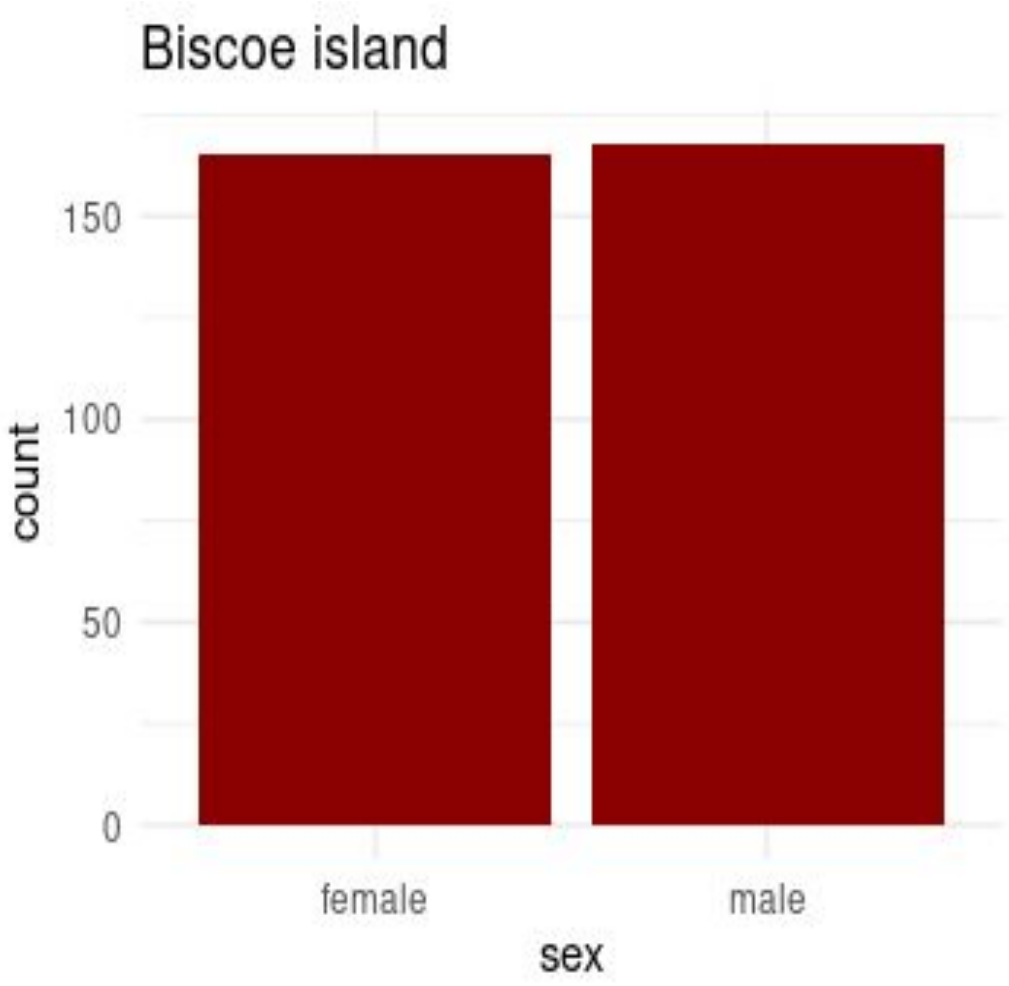
- thick line: **median**
- lower line: **25th percentile**
- upper line: **75th percentile**
- whiskers: **further non outlier points**
- points: **outliers**

4 Statistical tests for proportions and means

Statistical tests: one sample z-test for proportion

CIDRA-AIPD African Institute for Professional Development

Is the female to male ratio for penguins statistically similar to the expected one (0.5)?



prop.test tests if an observed proportion is equal to a certain expected value (z-test).

```
test_sex <- prop.test(
  x = 165, # number of successes (female)
  n = 333, # total number of trials (total num penguins)
  p = 0.5, # we test for prob = 0.5
  conf.level = 0.95 # confidence level
)
```

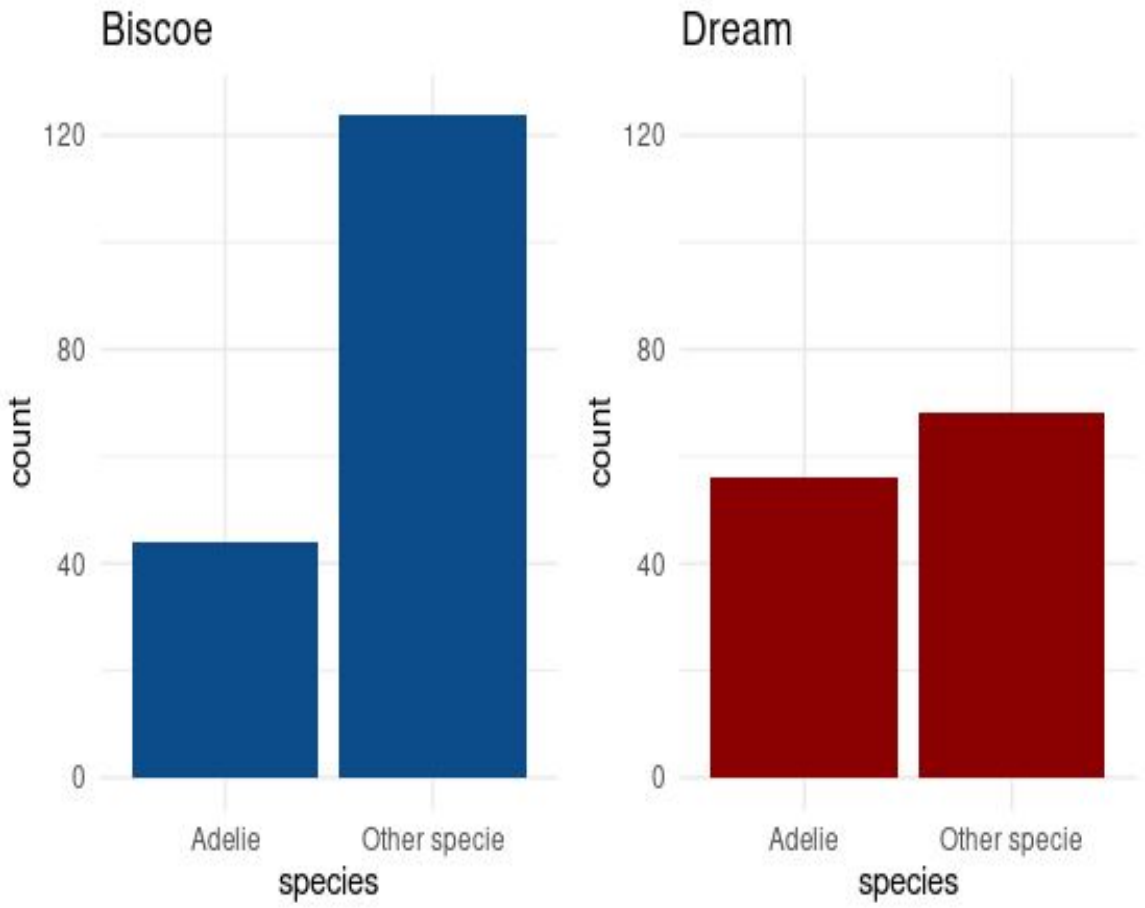
```
        1-sample proportions test with continuity correction

data:  165 out of 333, null probability 0.5
X-squared = 0.012012, df = 1, p-value = 0.9127
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4406707 0.5504259
```

p-value = 0.91 and the estimated proportion is [0.44,0.55] → fail to reject H0 of 50% female

Statistical tests: two sample z-test for proportion

CIDRA-AIPD African Institute for Professional Development

Are proportions between Adelie and other species statistically similar on different islands?



prop.test also tests if proportions are similar between two groups (z-test 2 sided)

```
#testing for equality of species on Biscoe vs Dream
adelie <- c(44, 56)
total_penguins <- c(168,124)

#p value << 0.05  we can reject H0 with high confidence
prop.test(adelie, total_penguins)
```

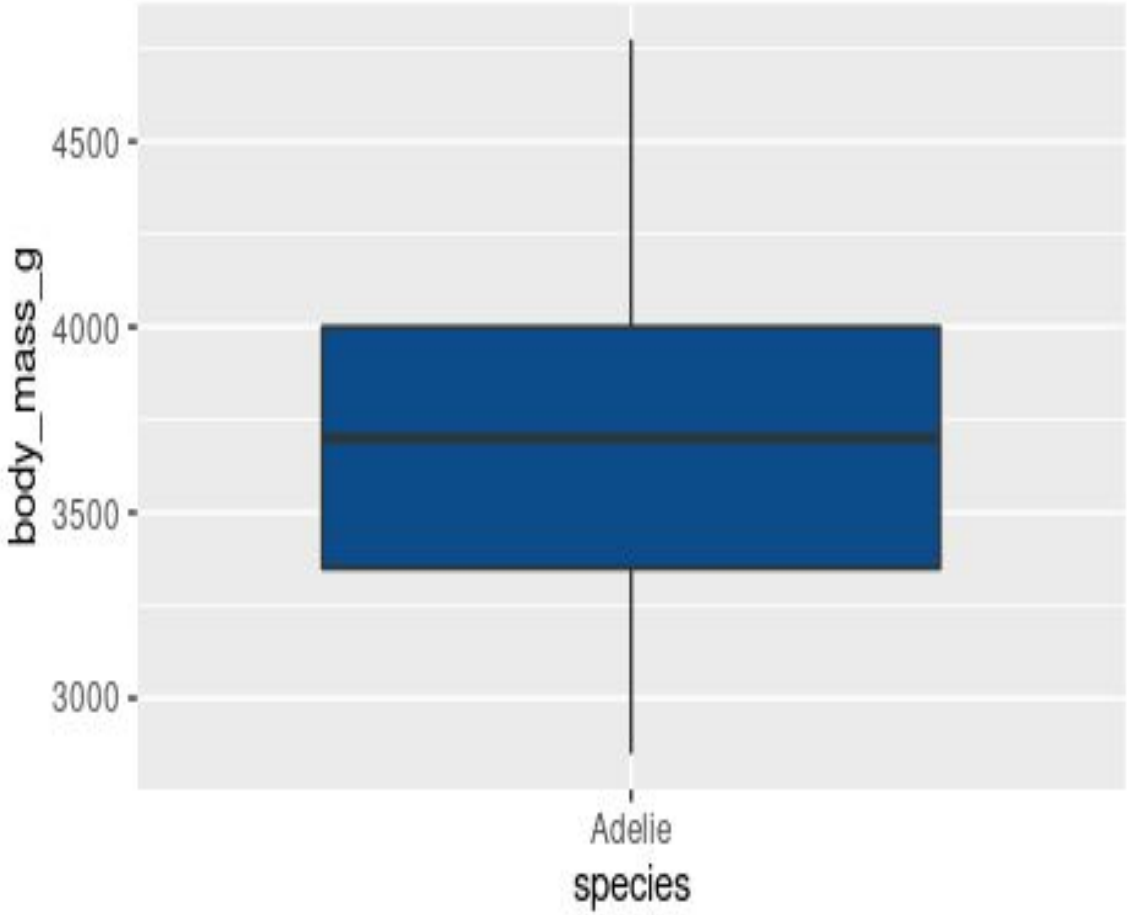
```
        2-sample test for equality of proportions with continuity correction

data:  adelie out of total_penguins
X-squared = 10.575, df = 1, p-value = 0.001146
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.30668273 -0.07273355
sample estimates:
   prop 1    prop 2
0.2619048 0.4516129
```

p-value = 0.0011 →

We can reject the null hypothesis (proportions are statistically different)

Statistical tests: one sample t-test for means

Is the average mass of Adelie equal to 3.6 kg?



t.test performs **one** and two sample t-tests on vectors of data

```
#One Sample t-test with almost correct mean, unknown variance
test_right <- t.test(dat$body_mass_g,
                     mu = 3600,
                     alternative = 'greater')
```

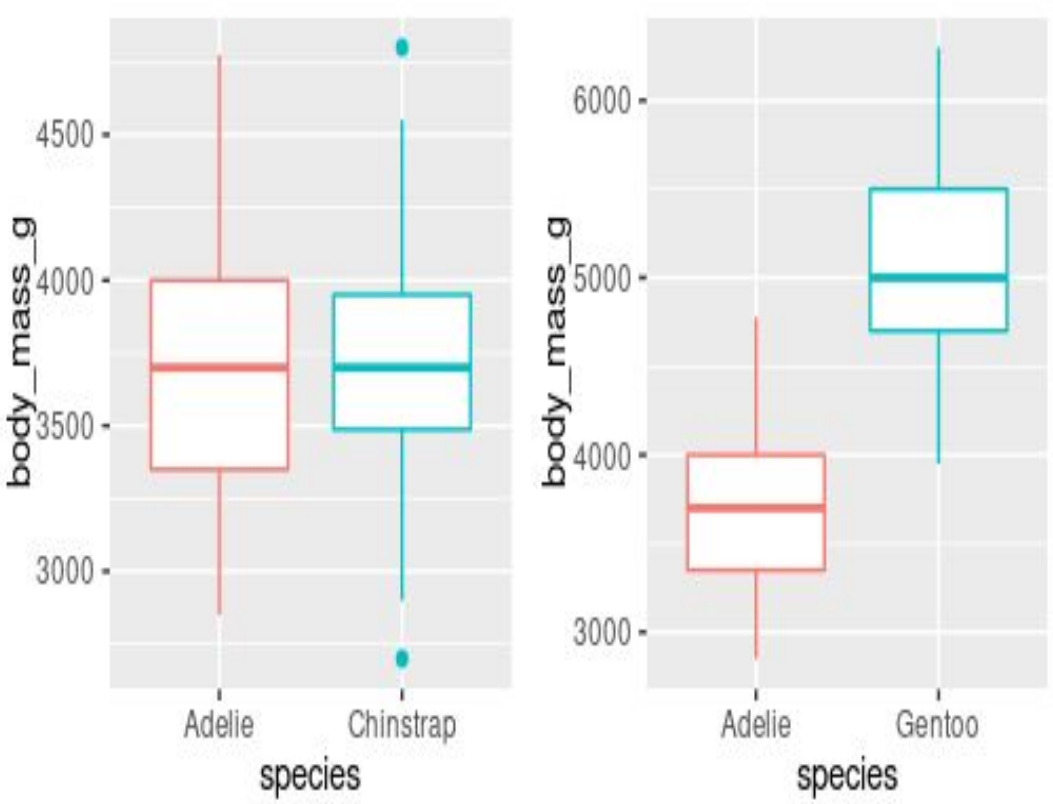
```
        One Sample t-test

data:  dat$body_mass_g
t = 2.6974, df = 150, p-value = 0.003894
alternative hypothesis: true mean is greater than 3600
95 percent confidence interval:
 3638.899      Inf
```

p-value = 0.004 →
We can reject the null hypothesis
(the mean is not stat. equal to 3.6kg)

Statistical tests: two sample t-test for means

CIDRA-AIPD African Institute for Professional Development

Is the average mass of Adelie penguins statistically different from those of Gentoo and Chinstrap?

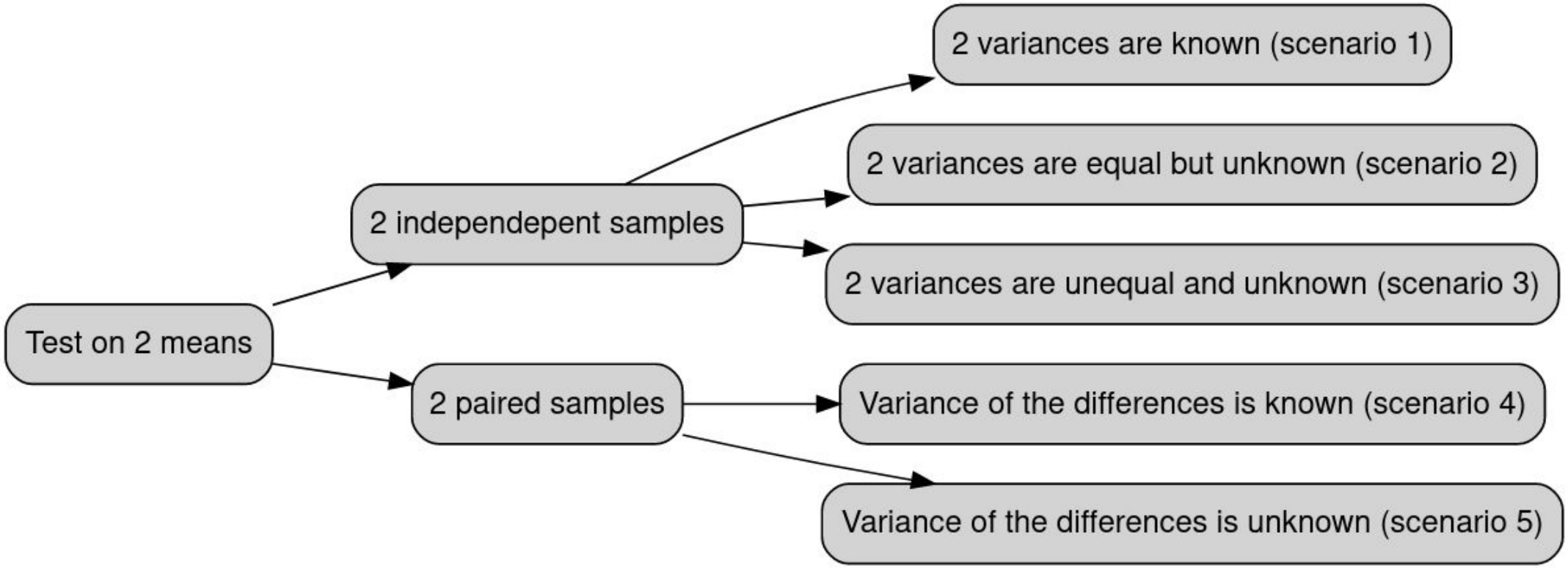


t.test performs one and **two** sample t-tests on vectors of data

```
#Welch Two Sample t-test 2 unequal and unknown variances
test_ad_gentoo <- t.test(body_mass_g ~ species,
                         data = dat_adelie_gentoo,
                         var.equal = FALSE,
                         alternative = "less"
)
> cat('p value adelie chinstrap = ', test_ad_chin$p.value)
p value adelie chinstrap =  0.2939304
> cat('p value adelie gentoo = ', test_ad_gentoo$p.value)
p value adelie gentoo =  3.854912e-65
```

We cannot H0 for Adelie vs Chinstrap
but we can reject H0 for Adelie vs Gentoo

Statistical tests: which t-test should I perform?



Source: **"Stats and R"** blog by **Antoine Soetewey** for this plot and a more general version of it

5 ANOVA

ANOVA: motivation and quick reminder

CIDRA-AIPD African Institute for Professional Development

The main aim of ANOVA is to compare **more than 2 groups** in **a statistically sound way**

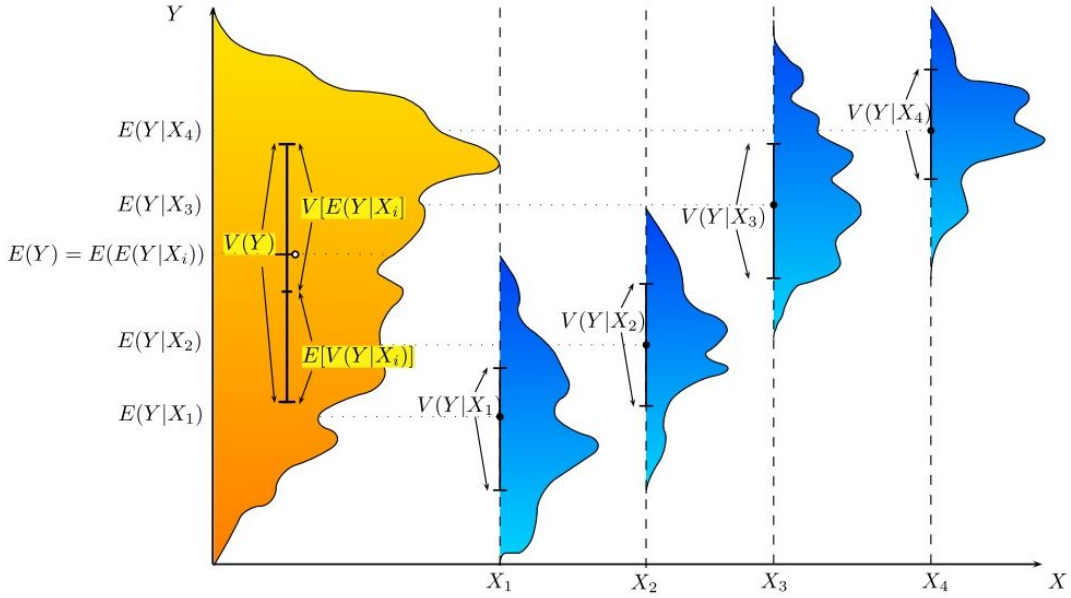
Probability of observing one significative results
due to chance for 3 groups

$$
\begin{aligned}
P(\text{at least 1 sig. result}) &= 1 - P(\text{no sig. results}) \\
&= 1 - (1 - 0.05)^3 \\
&= 0.142625
\end{aligned}
$$

ANOVA: analysis of variance

$$\frac{variance_{between}}{variance_{within}}$$

is larger than a certain threshold (5%)
groups are considered different



- Independence of observations
- Normality for the distributions of the residuals
- Equality (or "homogeneity") of variances

Source: Wikipedia

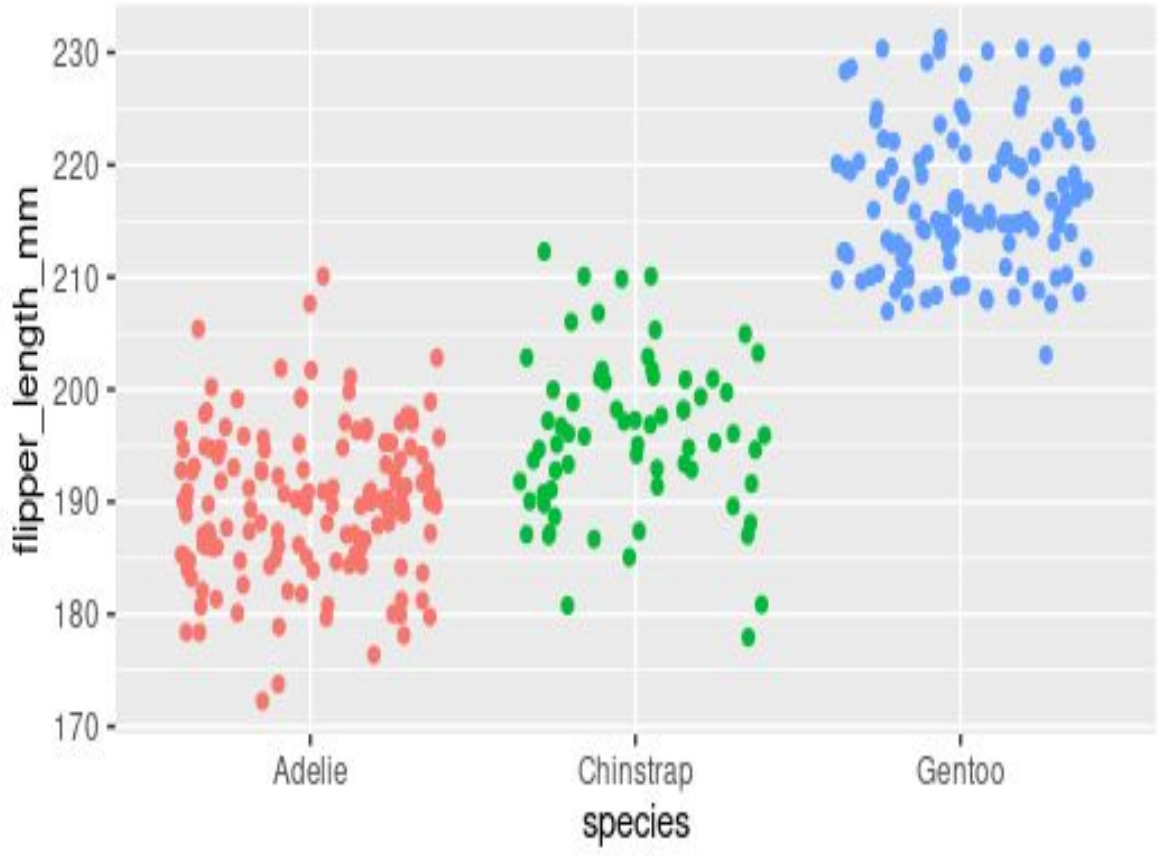
ANOVA: visual analysis

CIDRA-AIPD African Institute for Professional Development

Do the 3 species have statistically significant different flipper lengths?

Gentoo seem to have longer flippers!

```
#visualizing the flipper length per specie
ggplot(dat) +
  aes(x = species, y = flipper_length_mm, color = species) +
  geom_jitter() +
  theme(legend.position = "none")
```



geom_jitter is a shortcut for geom_point(position = "jitter").

It adds a **small amount of random variation** to the location of each point for visualization

ANOVA: perform ANOVA with R

CIDRA-AIPD African Institute for Professional Development

Do the 3 species have statistically significant different flipper lengths?

Test for Equal Means in a One-Way Layout

```
# 1st method for ANOVA
oneway.test(flipper_length_mm ~ species,
            data = dat,
            var.equal = TRUE # assuming equal variances
)
```

Fit an Analysis of Variance Model

```
# 2nd method for ANOVA (more info)
res_aov <- aov(flipper_length_mm ~ species,
               data = dat
)
```

```
> #ANOVA summary for this method. In this particular case
> # groups are sign. different since p is very small
> summary(res_aov)
             Df Sum Sq Mean Sq F value Pr(>F)
species       2  52473   26237   594.8 <2e-16 ***
Residuals   339  14953      44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2 observations deleted due to missingness
```

We reject the H0 that the 3 means are equal due to the very low Pr(>F) (i.e. p)

We do not know “how” (Post-hoc tests)

6 Conclusions

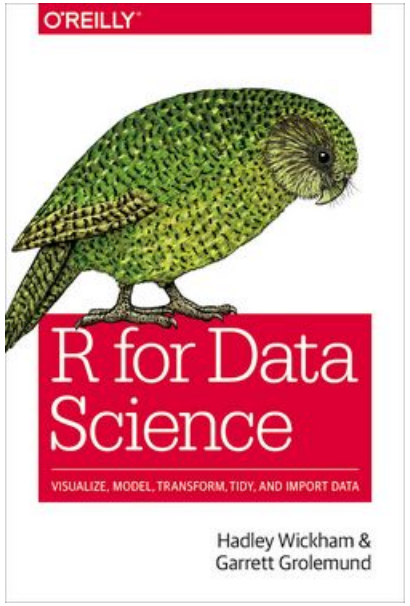
Conclusions

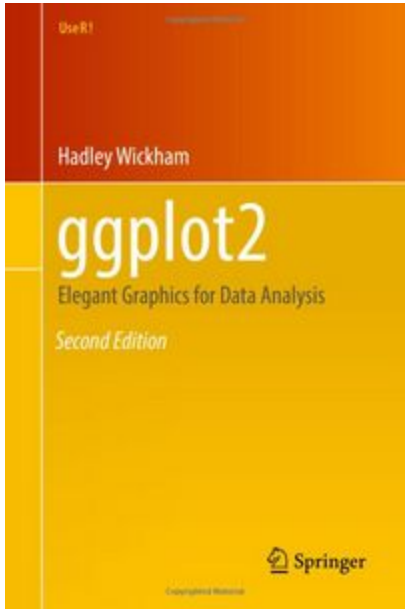
CIDRA-AIPD African Institute for Professional Development

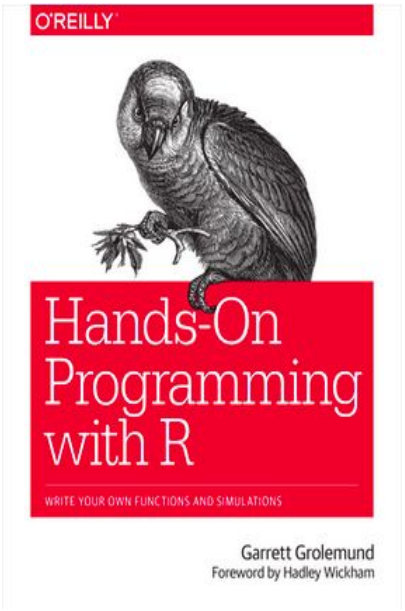
- R is a **powerful language for statistical computing**, easy to learn and with a vibrant community (R blog)
- **some of the basics of R**: installation, data types, operators, tibbles, …
- ggplot for visualization is **powerful but can get complex fast**
- statistical tests are **easy to perform in R, but be careful which one you choose!**
- Please check important topics **that we don't have the time to cover in this course (e.g. functions, loops, recycling rules, modelling, …)**

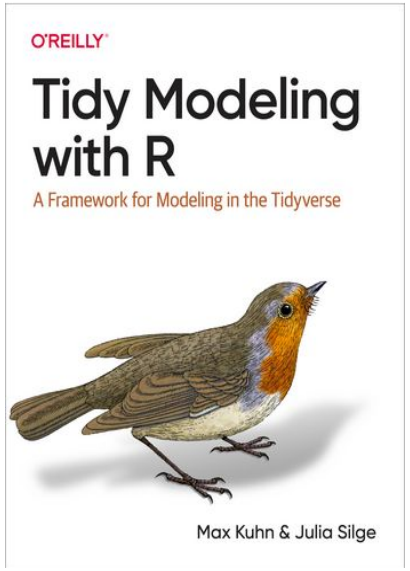
Additional material

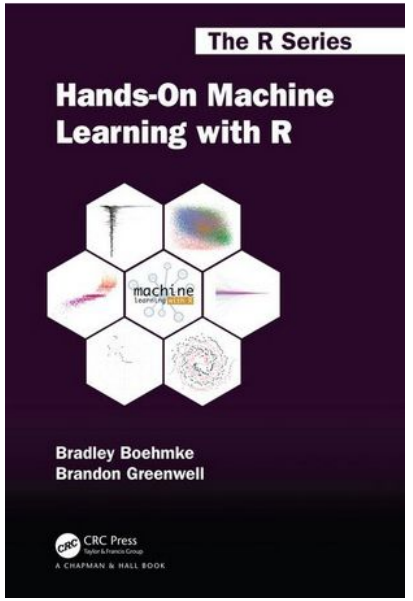
CIDRA-AIPD
African Institute
for Professional
Development











Contact me!

CIDRA-AIPD
African Institute
for Professional
Development

Github containing **slides and code** for this lecture

GitHub

Do not hesitate to contact me
if you have any further question about this lecture

luca.alb.rizzo@gmail.com

My LinkedIn profile

Gmail

Thank you